Communications and Protocols

Active Networks and Active Object Storage

John A. Chandy
Department of Electrical and Computer Engineering

Janardhan Singaraju, Ajith Thamarakuzhi, Cengiz Karakoyunlu, Orko Momin, Mike Runde, Paul Wortman

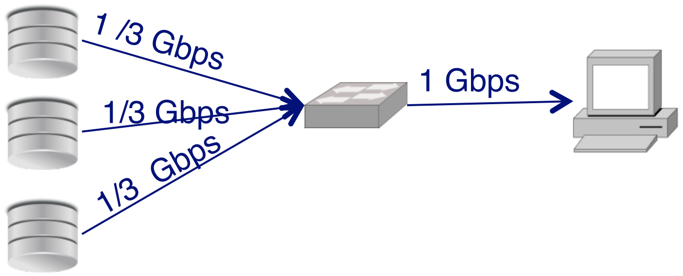
University of Connecticut

HEC FSIO Workshop
August 9, 2011

Active Storage Networks

- Active Disks
  - Intelligence at the disk can distribute computation to parallel disks
  - Process data in streams
  - Disks only have local view of data
- Active Storage Network
  - Network has a global view of data
  - Distributed caching of file system metadata and data
  - Redundancy optimizations

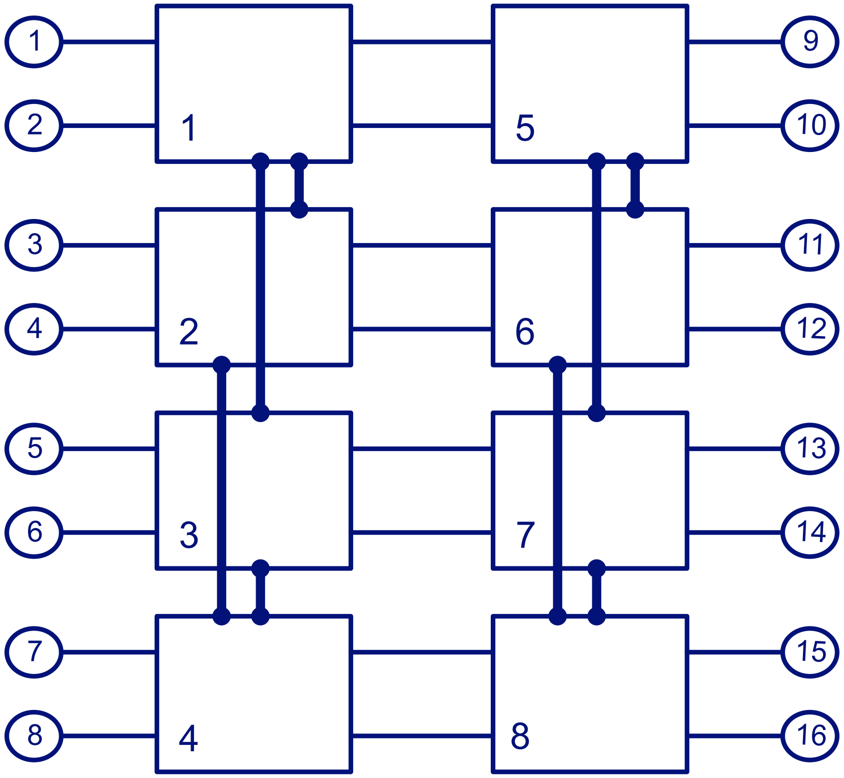
Active storage networks



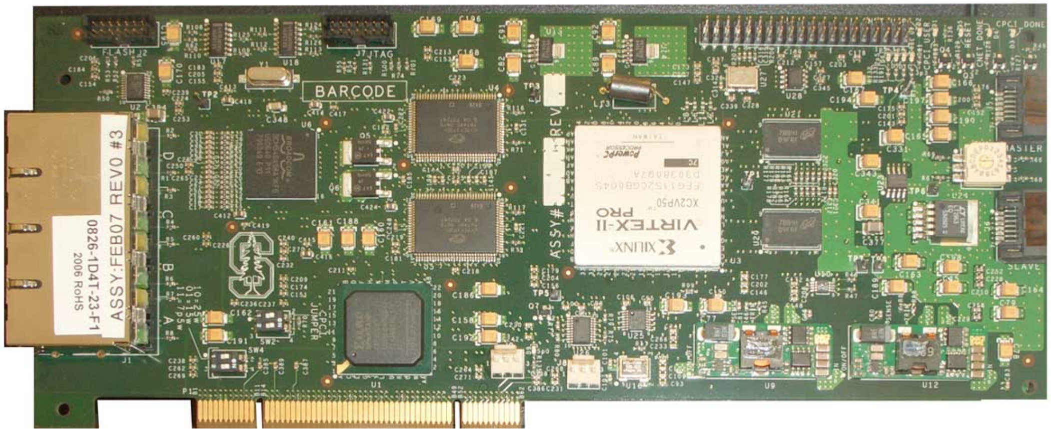
- An ASN is comprised of a smart switch along with intelligence embedded in the I/O network.
- Network Switches have global view of the data and can perform in-stream data reduction and transformation operations.
- ASN can enhance storage node performance as well as improve the computational performance of the parallel I/O systems.

Network switch topology

- 2-dilated flattened butterfly



Hardware Implementation



- NetFPGA board from Stanford
- 4 GigE connects
- 2 SATA connectors for node to node communication
- PCI bus for node to node communication

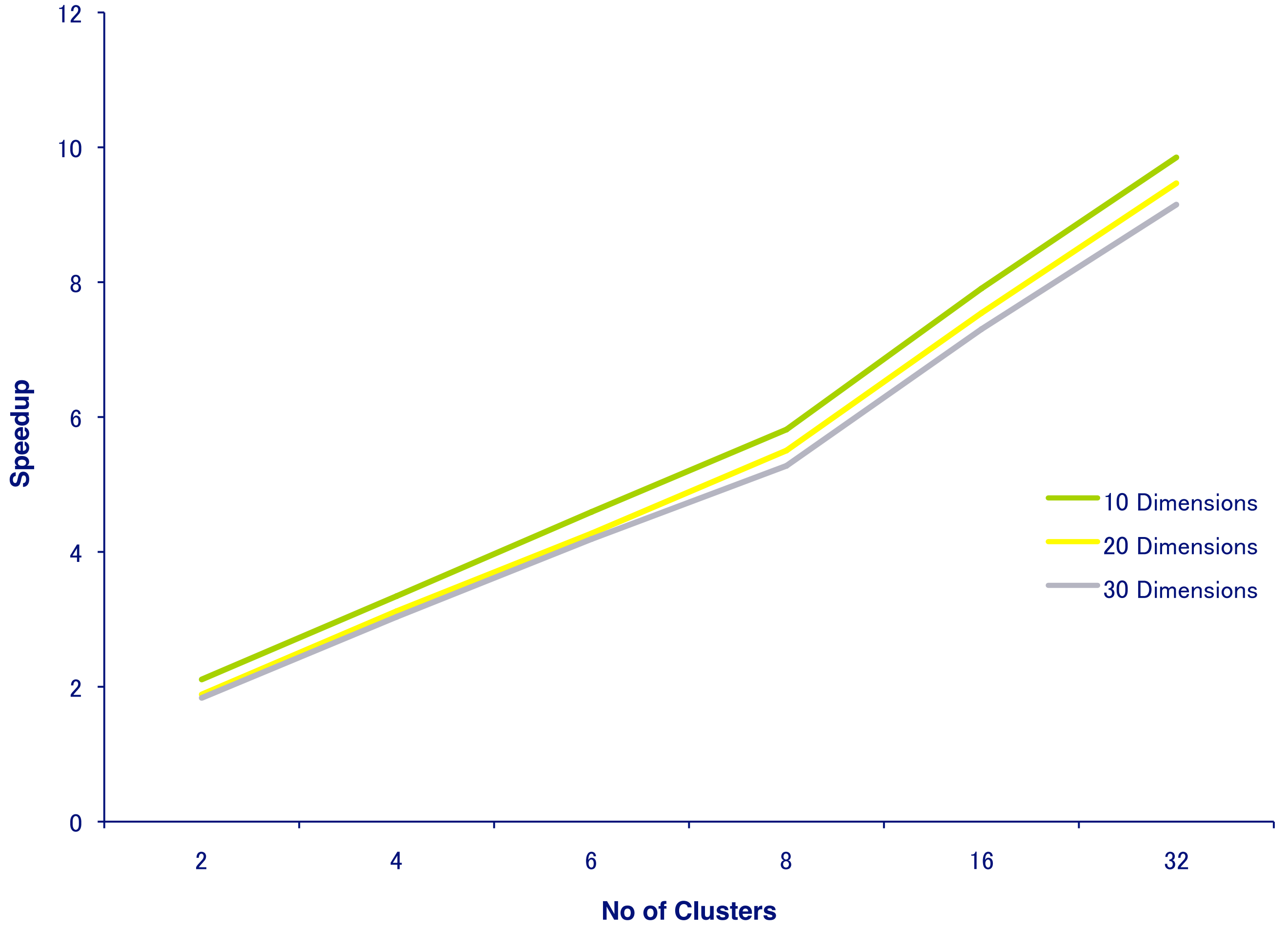
Active Storage Networks

- Application operations
  - – Reduction operations - min/max, k-means clustering, search
  - – Transformational operations – streaming, sort,
- File System Operations
  - – Locking
  - – Redundancy optimizations

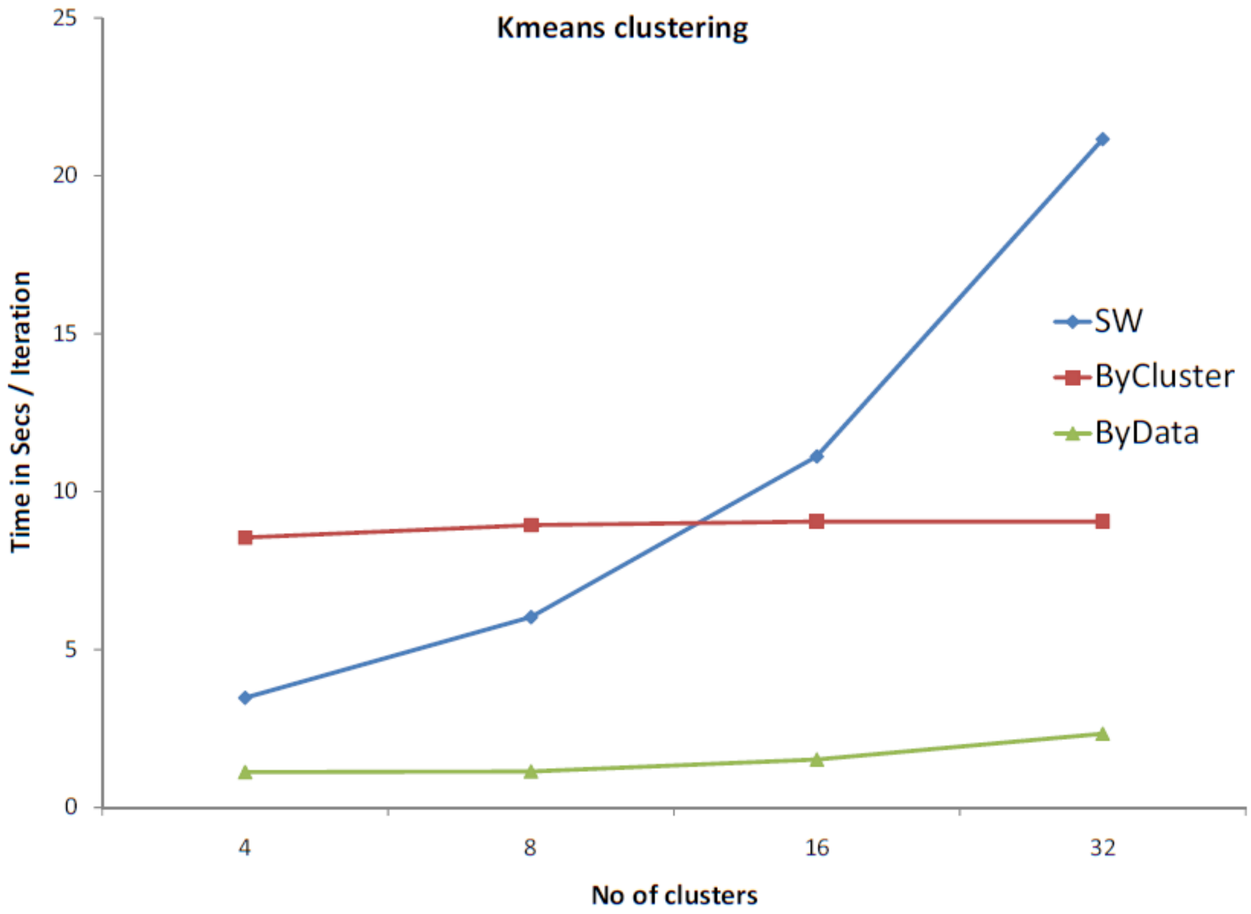
Parallelization techniques

- Functional units are re-used on reaching the reconfigurable hardware area limits.
- Data level parallelism by distributing the data to several functional units in several switch elements.
- Functional level parallelism by distributing functions to several elements.

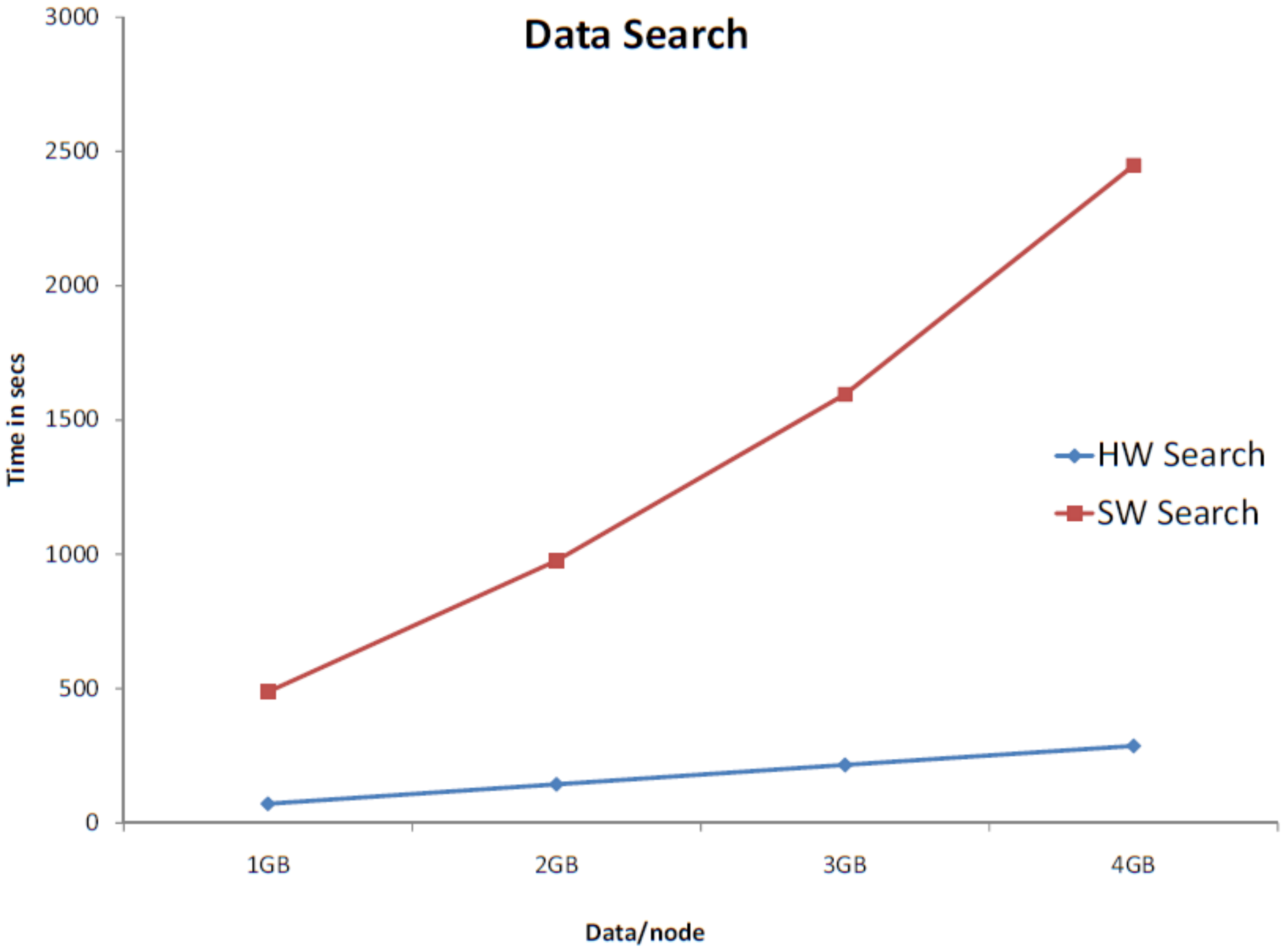
K-means clustering



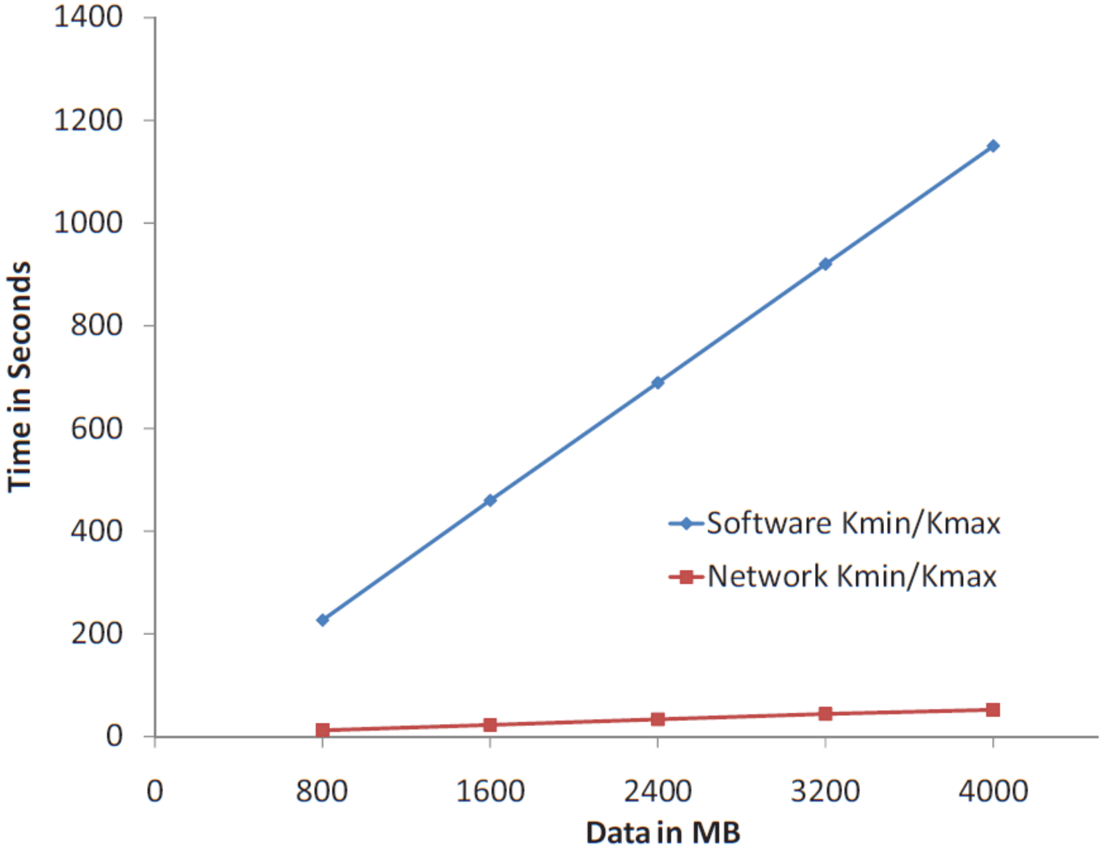
Runtime per iteration



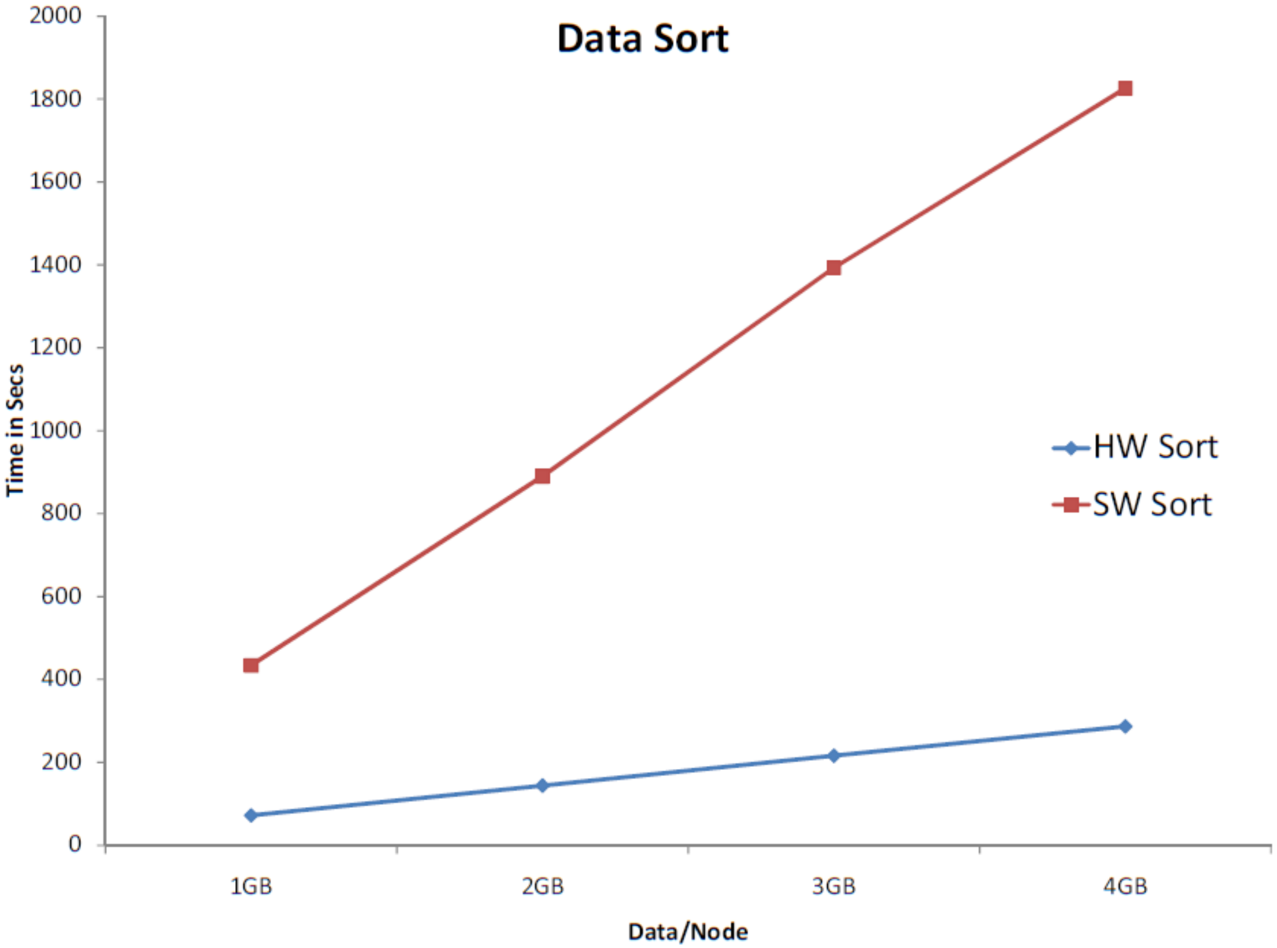
Data search


Data Search
Time in secs
3000
2500
2000
1500
1000
500
0
1GB
2GB
3GB
4GB
Data/node
HW Search
SW Search

Kmin/Kmax

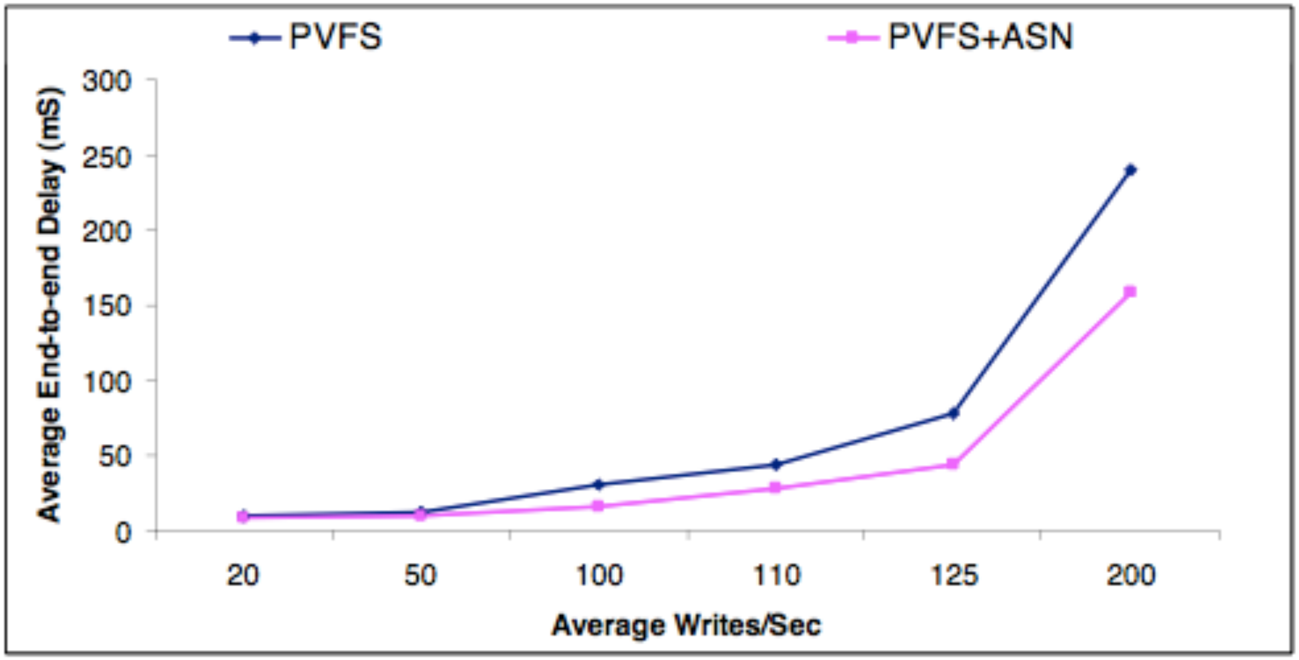

1400
1200
1000
800
600
400
200
0
Time in Seconds
0
800
1600
2400
3200
4000
Data in MB
Software Kmin/Kmax
Network Kmin/Kmax

Data sort


Data Sort
Time in Secs
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1GB
2GB
3GB
4GB
Data/Node
HW Sort
SW Sort

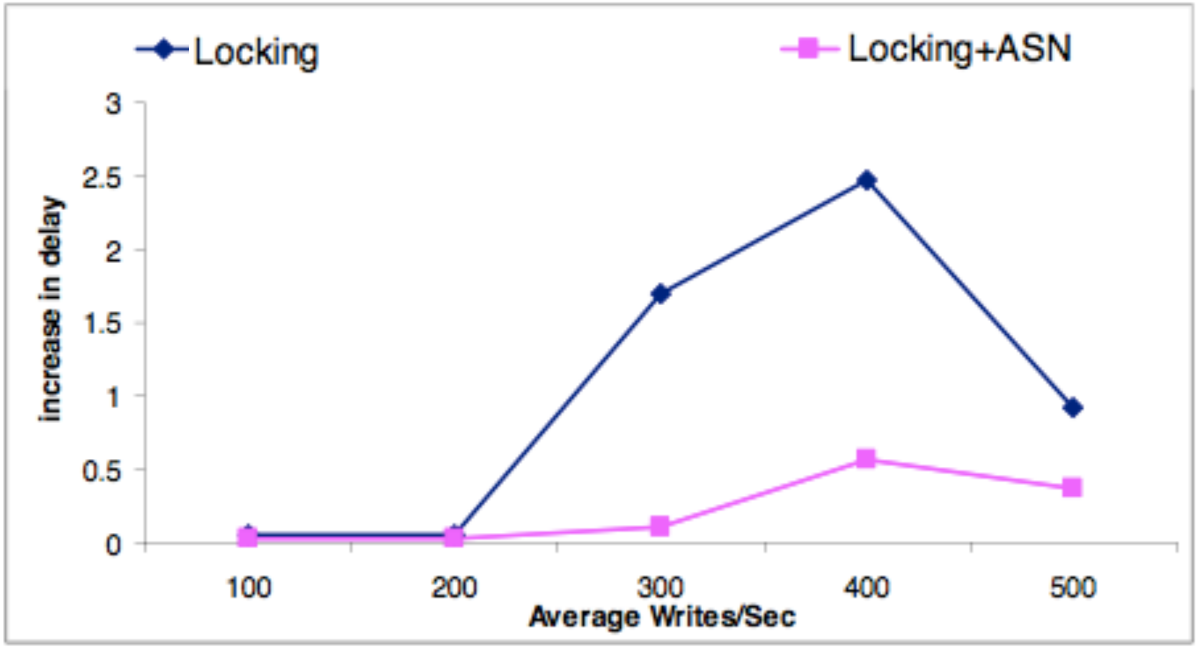
Redundancy optimizations

- RAIDed files, parity calculated in switch



File locking

- Lock table in switch



Active Storage Networks

- Lessons Learned
  - Hardware design is hard
  - HW Libraries can help
  - ASNs make most sense for reductions
  - Storage systems optimizations show promise
- What needs to get done
  - Better HW design
  - Application and FS hooks
  - When to do ASN and when to do SW?

Active Object Storage

- Active Disks
  - Intelligence at the disk can distribute computation to parallel disks
- Active Object Storage for Parallel File Systems
  - Active Disks for OSDs
  - Use Active Storage to improve parallel file system performance
  - Use Active Storage to improve parallel file system reliability
  - Application aware storage and autonomic storage using active OSDs.

Active Disks

- Can we use OSDs to make Active Disks a reality?
  - – Application-aware storage
    - Object attributes can give hints to the disk
    - Application specific
  - – Parallel File Systems
    - Felix et al. added a filtering layer to Lustre to provide active processing
  - – T10 OSD?

Active Disks using OSD

- Previous Implementation
  - – Based on disc-osd
  - – Object-oriented (Java)
    - Attach object types to storage objects
    - Define methods for object types
- New Implementation
  - – Based on osc-osd (supported by Panasas)
  - – RPC - Call functions on OSD remotely
  - – Execute Engines – C, Java, Python, etc.

Active Disks using OSD

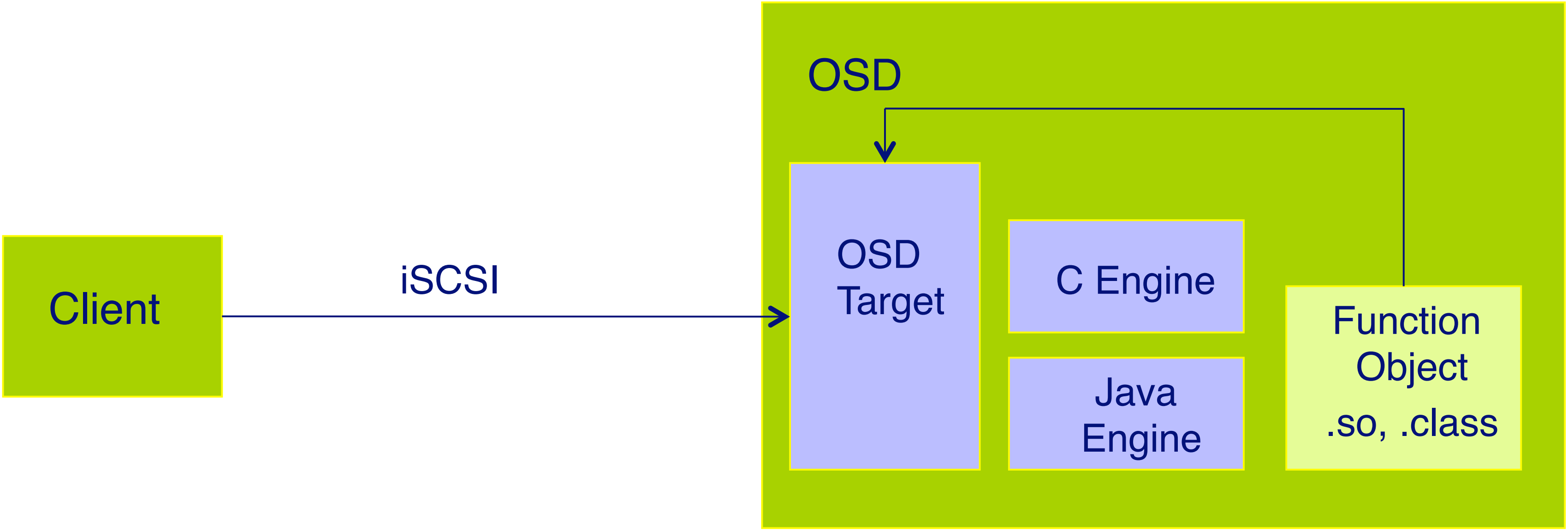
- How do you move code from client to target within OSD framework?
  - Create an object with the code
  - Each function object has a special attribute that defines the type of associated execute engine
  - OSD can support multiple execute engines

Active Disks using OSD

- How do you execute the method remotely within the OSD framework?
  - New EXECUTE FUNCTION command so that we can invoke a function
  - We use the CDB continuation to specify the parameters
  - Results (if any) returned directly or written to a new object

From T10/08-185r5 changes to OSD-2

Active Disks using OSD



Active Disks using OSD

- Status:
  - – C and Java engines complete
  - – Python engine soon
  - – OrangeFS support for OSDs

Summary

- Active storage networks
  - Improves performance of computation kernels
  - Useful in parallel file system optimizations
- Active storage for improved file system performance
- Acknowledgements: NSF CCF-0621448, CCF-093787

Communication and Protocols

- Coherence schemes
- Scalable abstractions for scientific data
- Scalable replication, relocation, failure detection, and fault tolerance
- Topology aware storage layout
- Wide area storage access protocols
- Cloud storage?
- Inter-stack communication?
- Memory hierarchy?